**Conner Olsen**

**CS539**

**Analysis Report**

**Optimizing AP Exam Selection for Maximizing College Credits**

**Overview:**

The objective of this project is to derive a machine-learning model-based solution that maps out the most efficient AP schedule from freshman to senior year, which would maximize potential college credits while ensuring the workload is manageable and prerequisites are met.

The research question is: Can a machine learning model, using a dataset of AP exams' mean scores, workload, and prerequisites, effectively cluster AP courses into categories that enable the creation of an optimal sequence for high school students to maximize their potential for earning college credits?

To answer this question, I used a dataset of AP exams' mean scores, workload, and prerequisites to train a gaussian mixture clustering model. The model was then used to cluster AP courses into categories that enable the creation of an optimal sequence for high school students to maximize their potential for earning college credits.

**Summary:**

This project focuses on optimizing AP course selection for high school students to maximize college credits. Utilizing machine learning, it clusters AP courses based on mean exam scores, workload, prerequisites, and course sequences. The aim is to devise optimal course pathways, balancing workload and prerequisites for maximized college credit acquisition.

**Project Description:**

The motivation behind this project stems from the observation that a well-planned AP course sequence can significantly influence a student's college readiness and credit accumulation. By leveraging machine learning, the project seeks to analyze AP courses based on workload, prerequisites, and exam success rates, providing students with data-driven recommendations for their high school academic journey.

I hoped to discover the following:

1. Identify an optimal sequence of AP courses.

2. Understand the balance between course workload, prerequisites, and exam success rates.

3. Provide data-driven recommendations for course selection.

The datasets used are the following:

1. College Board's Digital Services Dataset 2019:

• Description: This dataset provides detailed statistics on AP test scores across various subjects.

• Columns: ETHNICITY, MEAN SCORE, NUMBER OF STUDENTS FOR EACH EXAMINATION, NUMBER OF STUDENTS AT EACH LEVEL, TOTAL EXAMS, and breakdowns for each AP subject.

• Source: http://media.collegeboard.com/digitalServices/pdf/research/2013/National\_Summary\_13.xls

2. Martha's Vineyard Regional High School AP Workload Information Dataset:

• Description: This dataset provides a breakdown of the workload expectations for each AP class at Martha's Vineyard Regional High School.

• Columns: AP Course, Homework Expectations, Major Assessments, Other Notes

• Source: https://www.mvrhs.org/wp-content/uploads/2013/10/AP-Workload-Information.pdf

3. South County High School Expectations for Advanced Placement Courses Dataset:

• Description: This dataset provides a breakdown of the workload expectations for each AP class at South County High School.

• Columns: AP Course, amount of reading per week, "frequency of Tests, Essays/Times Writings, Papers etc", number of Major Projects, summer assignments.

• Source: https://southcountyhs.fcps.edu/sites/default/files/media/inline-files/Advanced%20Placement%20Course%20Expectations\_1920.pdf

4. LYME-OLD LYME HIGH SCHOOL Workload Expectations for Advanced Placement (AP) Dataset:

• Description: This dataset provides a breakdown of the workload expectations for each AP class at Lyme-Old Lyme High School.

• Columns: Summer work, Weekly homework time commitment.

• Source: https://resources.finalsite.net/images/v1591795986/region18/bagsrsrn76kdpfa3puo9/Workload\_Expectations\_for\_AP.pdf

**Methods:**

**Data Collection and Cleaning:** The project utilizes the College Board's Digital Services Dataset and high school workload datasets, requiring extensive cleaning and integration. This involved removing unnecessary information, extrapolating missing data, converting non-numeric data to numeric form, and combining data sources using AP course names as keys.

**Exploratory Data Analysis (EDA):** The EDA includes analysis of data distribution, identifying outliers, examining relationships among columns, and conducting statistical tests. Visualization techniques such as histograms, scatter plots, bar plots, line plots, and radar/spider plots were employed for a comprehensive understanding of the data.

**Feature Engineering:** Transformation of data to enhance model performance was undertaken. This involved converting raw data into a format suitable for analysis, like converting the number of assignments and tests into weekly time commitments.

**Modeling:** Various clustering algorithms were applied, including K-Means, DBSCAN, Hierarchical Clustering, Gaussian Mixture Models, Spectral Clustering, and Self-Organized Maps. Each model was carefully chosen and applied to the dataset to find optimal clustering of AP courses.

**Cluster Evaluation:** Post-modeling, clusters were evaluated using statistical tests and data visualization to assess the effectiveness of the clustering approach.

**Conclusion:**

I created a dataset, Final\_Result.xlsx, of AP courses, including workload, prerequisites, and exam success rates, and used it to train a gaussian mixture clustering model. The model was then used to cluster AP courses into categories that enable the creation of an optimal sequence for high school students to maximize their potential for earning college credits. The results of the clustering model were evaluated using statistical tests and data visualization to assess the effectiveness of the clustering approach. Using the results of the clustering model I created labeled the clusters and created a dataset of AP courses with recommended sequences.